SCIENCE, TECHNOLOGY & INNOVATION INDICATORS

Open indicators: innovation, participation and actor-based STI indicators
PARIS 2017

Topic modelling approaches to aggregated citation data[1]

Johan Eklund*, Gustaf Nelhans*

*johan.eklund@hb.se; gustaf.nelhans@hb.se
University of Borås, Swedish School of Library and Information Science, SE 501 90 Borås, Sweden

Abstract
In this research in progress paper we report on preliminary results from the proposed novel uses of topic modelling approaches to bibliographic references as sources for "bags-of-words" instead of actual text content in scientometric settings. The actual cited references, viewed as concept symbols for paradigmatic approaches to earlier research, could thereby be used to cluster research. We will demonstrate an explorative approach to using cited reference topics for the discovery of hidden semantic reference structures in a set of scientific articles. If found fruitful and robust, this approach could complement existing text based and citation based techniques to clustering of research that might bridge the two approaches. By approaching references as "words" and reference lists as "sentences" (or documents) of such "words", we demonstrate that the topical structure of document collections can also be analyzed using an alternative and complementary source of content, which additionally provides an interesting perspective on bibliographic references as units of a meta language describing document content.

INTRODUCTION
Scientometric analyses has traditionally mainly focused on the links between documents based on citations and especially the network model of the structure of publications (Price, 1965), and the development of aggregated citation metrics based on bibliometric coupling and co citation analysis at the article level for identifying 'scientific specialties' (Griffith, Small, Stonehill, & Dey, 1974; Kessler, 1963; Small & Griffith, 1974). Co-citation analysis was subsequently developed empirically and theoretically on higher levels of aggregation, into focussing on co-citation at the authorship level (McCain, 1986; White & Griffith, 1981); and later at the journal level (McCain, 1991a, 1991b).

At the same time text-based analysis on the contents of scholarly texts such as publications has not seen the same amount of development within scientometrics. The most notable attempts were developed at the border between scientometrics and STS in the early 1980's under the umbrella term co-word analysis. (Michel Callon, Courtial, Turner, & Bauin, 1983; M. Callon, Courtial, & Laville, 1991; Michel Callon, Law, & Rip, 1986). Instead of looking at the relationship between documents based on citation links, the researchers grouped research together based on shared textual content of the research in question. This strand, although discussed within the bibliometrics community was somewhat debated (Courtial, 1998; Leydesdorff, 1997, 1998), has found new use in present day software (e.g. *VOSviewer*,

[1] This work was supported by Minso Solutions AB, Sweden

developed at Universiteit Leiden), partly due to the increasingly powerful capabilities of computers and the increasing availability of full text scientific publications for possible use in the analysis (van Eck & Waltman, 2011).

Lately, new methods of analyzing large sets of texts have been developed under the heading of Topic modelling. (Blei, Ng, & Jordan, 2003; Jockers, 2014) Topic modelling is a statistical approach to exploratory text analysis, with the aim to identify latent factors, called "topics", that can be used to probabilistically explain and group words (or possibly *n*-grams, see e.g. Wallach, 2006). However, these methods are not limited to the analysis of ordinary natural language texts – as we demonstrate in this paper, topic modelling is also a useful tool for the exploration of other units of content, such as reference lists. This opens a new avenue in the indexing and clustering of research based on publication texts.

In Scientometrics, topic modelling techniques have been used to study a wide array of different purposes ranging from representations of topics within a research field and thematically labelling authors based on contents of research (Kongthon, Haruechaiyasak, & Thaiprayoon, 2008; Lu & Wolfram, 2012) to dynamic analysis of dynamic patterns and emerging trends within research fields (Jensen, Liu, Yu, & Milojevic, 2016) Comparison of topic modelling approaches and traditional citation based metrics and/or co-word analysis have also been performed and the results as of now are promising but still inconclusive about their effectiveness compared to traditional methods. (Loet Leydesdorff & Nerghes, 2017).

A specifically interesting theme of research approaches the combination of text based topic modelling techniques approaches with scientometric link based citation analyses. Two specific studies have been identified that may represent this category. First we find Erjia Yan's paper, proposing a topic-based PageRank algorithm combines citation analysis and the PageRank algorithm, commonly used in the Google web search algorithm (Yan, 2014) and Hassan and Naddawy's paper that combines topic modelling approaches with the identification of papers that cite each of these topics in a reciprocal mode (Hassan & Haddawy, 2015)

While topic modelling or associated technologies using algorithms such as latent Dirichlet allocation (LDA) have been used and further developed for scientometric purposes, in the studies we have found the methods have only been developed using a bag-of-words approach, based on terms used in the texts, or assigned descriptive keywords of various forms of scholarly publications.

In this paper, we will direct our attention to the actual scientometric citation data and treat bibliographic references as the common denominator of study. Thus, we will apply topic modelling analyses to the bibliographies of scholarly publications in a way that instead of actual text, we will regard the actual references in the reference list as "words", and the larger unit (the reference list) as a "sentence" that can be transformed into a bag of words representation.

The conceptual idea is derived from the influential work of Henry Small and Eugene Garfield about the citation as a concept symbol that stands for something more than just the link between two documents (e.g. Small, 1978). Thus, instead of regarding cited references as simply the sources for individual notions in specific instances of the previous literature, it is viewed in a wider setting of paradigmatic exemplars in the literature that the cited reference is part of. By elucidating these interconnections, thematic areas could be identified and delimited.

Traditional co-word metrics utilize one way of describing this relationship – the pair-wise occurrence of mutually cited papers. In the present study, it is argued that there might be other fruitful relationships between the cited documents that could be elucidated using topic modelling algorithms. Instead of simply identifying the co-occurrence of two documents at the time, we could use the strength of the co-occurrence of references in citing documents to better relate documents to each other. This novel approach to exploratory content analysis builds on the advancements that have been made in the research area of distributional semantics, in which word co-occurrence statistics is employed to build models of semantic relatedness between words, for instance by embedding words in a vector space. Prominent examples of such approaches are latent semantic analysis (Deerwester et al., 1990), probabilistic latent semantic analysis (Hofmann, 1999), random indexing (Kanerva et al., 2000), and more recently the skip-gram and the continuous bag of words models of the word2vec algorithm (Mikolov et al., 2013), as well as GloVe (Pennington et al., 2014).

Material and methods

The source data set originated from a randomized set of 100 cited references from an original set of 17.185 references found in Swedish clinical guideline documents (published between 2010 and 2015 on a number of different subjects within clinical medicine). The set of references was used as a seed to develop the material treated in the analysis. In Web of Science, all identifiable documents that cited these 100 papers (n=8.108) were collected and used in the study, meaning that all citing documents contain at least one reference each to the seed material. The choice of material was done to ensure that there were some association between the documents used in the study. The references from these citing documents were treated as bags of words "references") in the study, meaning that `cited reference 1` was treated as one "word", `cited reference 2` another word, and so on. If a reference was found in multiple citing documents, it was given the same ID, so that each unique cited document in the set became one unique term in the vocabulary created. Then, a traditional LDA algorithm was run on the set of words/references at the sentence level, where each reference list was treated as one sentence.

The references in each article in the dataset have been encoded with standardized labels so that each list of references can be translated into a sequence of reference labels and used as a document content representation, for instance:

d_3: `ref10 ref11 ref5 ref6 ref1 ref12 ref13 ref14 ref15 ref8 ref16 ref9`

The resulting structure of references can be regarded as sentences of "words" (each label constituting a word in the "vocabulary" of references). For the topical analysis of these sentences we have applied the *latent Dirichlet allocation* (LDA) method (see Blei et al., 2013), which implements a probabilistic approach to topic detection in a dataset. Let D denote a set of documents and V a vocabulary of labels. Assuming that each reference appears at most once in the reference list of a document d_j we can model an arbitrary reference list as a subset of V. We further assume that there for each document is a set T of (latent) topics, distributed among the documents in D according to a multinomial distribution parametrized by a parameter vector $\boldsymbol{\theta}$. The parameter vector $\boldsymbol{\theta}$ is in turn assumed to be sampled from a Dirichlet distribution parametrized by a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, such that $k = |T|$ and $\alpha_i > 0$ for all indices. Furthermore, each word w_i in V is assumed to be conditionally dependent on a topic z_t according to a multinomial distribution parametrized by a vector $\boldsymbol{\beta}_t = (\beta_{t,1}, \ldots, \beta_{tn})$ such that $0 \leq \beta_{ti} \leq 1$ for all indices. Given these definitions, the probability of a document d, modelled as a sequence $\mathbf{w} = (w_1, \ldots, w_n)$ of n words, is given by (Blei et al., 2013):

$$p(\mathbf{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{i=1}^{n} \sum_{t=1}^{k} p(z_t|\boldsymbol{\theta}) p(w_i|z_t, \beta_{ti}) \right) d\boldsymbol{\theta}$$

The parameters **α** and **β** are obtained using maximum likelihood estimation. For topic modelling of references, given a topic z_t, we have selected the references w_i maximizing the probability $p(w_i|z_t, \beta_{ti})$.

Analysis

Preliminary results from the analysis using twenty topics shows promising results in terms of matching of corresponding references from the set of 8.108 citing documents. An example of the results obtained using this approach is shown in figure 1, in which we can observe a high estimated relative frequency for articles published in journals on periodontology as indicated by the source titles of the identified references. In the diagram, the blue bar represents the overall frequency of words (references) in the set and the red bar shows the estimated term frequency within the selected topic. This indicates that that a very large share of the references in the full set identified in this topic are in fact associated with the topic. The results obtained so far strongly indicate that topic modelling on references is a useful tool for exploring the topical structure of a document collection.

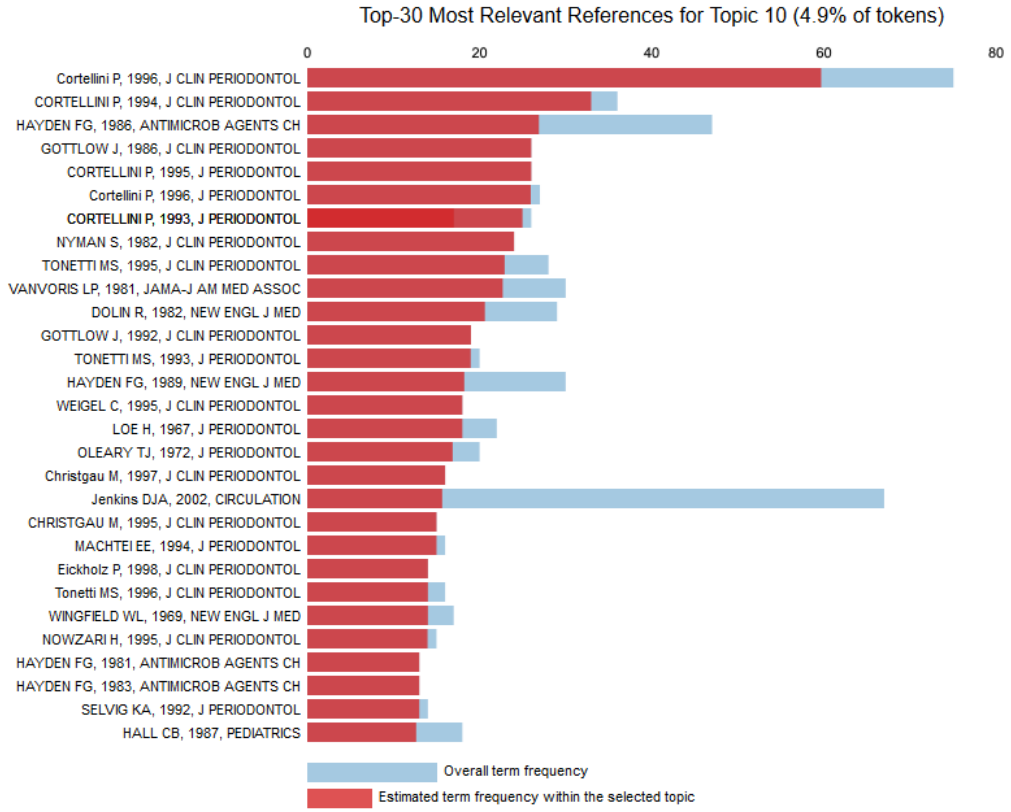

Figure 1. A topic containing a high density of articles on periodontology.

Discussion and conclusion

In the study, we used the citing side of documents to construct our topic models, but there is nothing that prevents turning into the other direction to treat the papers that cite the seed set of references the same way. Further, for practical reasons, only the cited references as found within

the reference lists of Web of Science for each of the papers were used as content in the analysis. These could be matched to the actual cited documents and all information from the actual entry in the citation index, such as actual text from the titles, abstract or keywords, as well as, matched information from other databases, such as MESH terms could be used in the visual representation of the topics for explorative purposes.

The present method of approaching cited references as text for topic modeling purposes shows promise to construct novel ways of combining text based information science approaches with established scientometric methods. As such, this approach could complement existing text based and citation based techniques to clustering of research that might bridge the two approaches. By approaching references as "words" and reference lists as "sentences" (or documents) of such "words", we demonstrate that the topical structure of document collections can also be analyzed using an alternative and complementary source of content, which additionally provides an interesting perspective on bibliographic references as units of a metalanguage describing document content.

References

Blei, D. N., Ng, A. Y., & Jordan, M. J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191-235. doi:10.1177/053901883022002003

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155-205.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the dynamics of science and technology: sociology of science in the real world*. Basingstoke: Macmillan.

Courtial, J. P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science, 49*(1), 98.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). The Structure of Scientific Literatures II: Toward a Macrostructure and Microstructure for Science. *Science Studies, 4*(4), 339-365. doi:10.1177/030631277400400402

Hassan, S. U., & Haddawy, P. (2015). Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy. *Scientometrics, 103*(1), 33-46. doi:10.1007/s11192-015-1528-3

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI'99)*, Kathryn B. Laskey and Henri Prade (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 289-296.

Jensen, S., Liu, X. Z., Yu, Y. Y., & Milojevic, S. (2016). Generation of topic evolution trees from heterogeneous bibliographic networks. *Journal of Informetrics, 10*(2), 606-621. doi:10.1016/j.joi.2016.04.002

Jockers, M. (2014). *Text Analysis with R for Students of Literature*: Springer International Publishing.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In Gleitman, L.R. and Josh, A.K. (Eds.): *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, p. 1036. Mahwah, New Jersey: Erlbaum, 2000.

Kessler, M. M. (1963). Bibliographic Coupling Between Scientific Papers. *American Documentation, 14*(1), 10-25. doi:10.1002/asi.5090140103

Kongthon, A., Haruechaiyasak, C., & Thaiprayoon, S. (2008). Enhancing the Literature Review Using Author-Topic Profiling. In G. Buchanan, M. Masoodian, & S. J. Cunningham (Eds.), *Digital Libraries: Universal and Ubiquitous Access to Information, Proceedings* (Vol. 5362, pp. 335-338).

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science, 48*(5), 418-427.

Leydesdorff, L. (1998). Reply about using co-words. *Journal of the American Society for Information Science, 49*(1), 98-99.

Leydesdorff, L., & Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the Association for Information Science and Technology, 68*(4), 1024-1035. doi:10.1002/asi.23740

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology, 63*(10), 1973-1986. doi:10.1002/asi.22628

McCain, K. W. (1986). Cocited author mapping as a valid representation of intellectual structure. *Journal of the American Society for Information Science, 37*(3), 111-122. doi:10.1002/(sici)1097-4571(198605)37:3<111::aid-asi2>3.0.co;2-d

McCain, K. W. (1991a). Core Journal Networks and Cocitation Maps - New Bibliometric Tools for Serials Research and Management. *Library Quarterly, 61*(3), 311-336.

McCain, K. W. (1991b). Mapping Economics Through the Journal Literature - An Experiment in Journal Cocitation Analysis. *Journal of the American Society for Information Science, 42*(4), 290-296. doi:10.1002/(sici)1097-4571(199105)42:4<290::aid-asi5>3.0.co;2-9

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates Inc., USA, 3111-3119.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Price, D. J. d. S. (1965). Networks of Scientific Papers. *Science, 149*(3683), 510-515.

Small, H. G. (1978). Cited Documents as Concept Symbols. *Social Studies of Science, 8*(3), 327-340.

Small, H. G., & Griffith, B. C. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies, 4*(1), 17-40.

van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *ISSI Newsletter, 7*(3), 50-54.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (ICML '06). ACM, New York, NY, USA, 977-984.

White, H. D., & Griffith, B. C. (1981). Author Cocitation – A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science, 32*(3), 163-171. doi:10.1002/asi.4630320302

Yan, E. J. (2014). Topic-based Pagerank: toward a topic-level scientific evaluation. *Scientometrics, 100*(2), 407-437. doi:10.1007/s11192-014-1308-5